Report for the Project "Profiling Fraud Victims Through Bank Transaction Data to Develop Enhanced Fraud Early-Warning Models"

Zongxiao Wu and Yizhe Dong*

University of Edinburgh

1. Introduction

The expansion of online banking, mobile payments, e-commerce, and other forms of digital finance has substantially increased the volume and speed of financial transactions. While these developments have improved the accessibility and efficiency of financial services, they have also created new opportunities for fraudulent activity (Motie and Raahemi, 2024). Financial institutions must identify suspicious behaviour within transaction streams that are too large and complex to be monitored manually. Fraud can generate direct financial losses, reimbursement and investigation costs, regulatory exposure, service disruption, and reputational damage. It may also undermine customers' confidence in digital financial services, particularly when vulnerable individuals are repeatedly targeted (Fanai and Abbasimehr, 2023).

Given the information available at a particular point in time, the objective of fraud detection is to assess whether a transaction is likely to be fraudulent or whether a customer's recent behaviour indicates an increased vulnerability to fraud. This problem is difficult for three main reasons. First, fraudulent transactions typically account for only a small proportion of the transaction population. A model may therefore achieve high overall accuracy while failing to identify the relatively small number of cases that matter most (Boulieris et al., 2024). Second, fraud strategies evolve over time, which limits the effectiveness of fixed rules and manually specified thresholds (Forough and Momtazi, 2021). Third, classification errors have asymmetric consequences. Failing to identify fraud may result in substantial financial loss, whereas incorrectly blocking a legitimate transaction may inconvenience customers and generate additional administrative costs (Fanai and Abbasimehr, 2023).

Traditional fraud-detection systems rely primarily on expert rules and predefined thresholds. These systems are transparent and straightforward to implement, but their effectiveness depends on whether the underlying rules remain relevant as fraud patterns change. Conventional machine learning methods provide greater flexibility, although they may not fully capture complex nonlinear relationships among transaction and customer characteristics. Deep learning methods can learn such relationships more effectively, but their predictions are often difficult to interpret. The choice of methodology therefore involves a trade-off between predictive flexibility and transparency.

This project examines whether bank transaction data can be used to support two related applications. The first is fraud-victim profiling, which seeks to identify behavioural, financial, and transaction-level

*The authors would like to express their sincere gratitude to the Smart Data Foundry and the Edinburgh Centre for Financial Innovations for their financial support of this project.

characteristics associated with greater vulnerability to fraud. The second is fraud early warning, which seeks to identify customers or transactions that may require timely intervention. Although these applications are closely related, they involve different prediction targets. The precise empirical design therefore depends on whether the available data contain customer-level victim labels, transaction-level fraud labels, or both. The broader methodological framework is based on Deep Boosting Decision Trees and combines the structured decision process of tree-based models with the representation-learning capacity of neural networks. We have conducted validation experiments on the open-source BankSim dataset using traditional machine-learning models to detect fraudulent transactions. The resulting customer profiles and risk scores can inform targeted interventions, including personalised warnings, adaptive transaction controls, and enhanced monitoring.

2. Methodology

The methodology is based on Deep Boosting Decision Trees (Xu et al., 2023). The central idea is to combine the interpretability of decision trees with the capacity of neural networks to learn nonlinear relationships. The framework has three main components: soft decision trees, a boosting structure, and a training strategy designed for highly imbalanced data.

The first component is a soft decision tree. In a conventional decision tree, each internal node applies a fixed rule and assigns an observation entirely to either the left or the right branch. A soft decision tree replaces this hard assignment with a probabilistic decision. Each internal node contains a small neural network that maps the input characteristics into the probability of following each branch. The leaf nodes produce customer- or transaction-level risk predictions. The prediction from a soft decision tree is obtained by combining the outputs of the leaf nodes, weighted by the estimated probability that the observation reaches each leaf. This structure allows the model to capture nonlinear relationships and interactions among the input characteristics while preserving a hierarchical decision process that can be examined.

The second component is boosting. A single soft decision tree may not capture all relevant patterns in the data. The framework therefore combines multiple soft decision trees, with each additional tree focusing on patterns that have not been adequately captured by the preceding trees. Earlier trees identify broad differences between legitimate and suspicious behaviour, while later trees refine the predictions for more difficult or ambiguous observations. The outputs of the individual trees are then aggregated to produce a final risk score.

The third component addresses class imbalance. Because fraud cases are rare, a conventional classification objective may be dominated by legitimate observations. A model trained to maximise overall classification accuracy may consequently perform poorly on the minority class. The framework therefore employs a compositional AUC-maximisation strategy rather than relying exclusively on classification accuracy or extensive over-sampling and under-sampling.

The training procedure has two related objectives. The first is to learn informative representations of transaction and customer characteristics. The second is to improve the model's ability to rank fraudulent or high-risk observations above legitimate or lower-risk observations. This ranking interpretation is particularly relevant in operational settings, where financial institutions may not have sufficient resources to investigate every transaction flagged by a model. A useful early-warning system should therefore place the cases most likely to require intervention near the top of the risk ranking.

Depending on data availability, the framework incorporates information such as transaction amount, transaction timing, merchant or payment category, recent transaction history, changes in customer behaviour, and customer-level profile indicators. Customer-history variables capture deviations from an individual's established behaviour, including unusual changes in transaction frequency, amounts, locations, or payment categories. The empirical design includes comparisons with established benchmark methods, including logistic regression, random forest, XGBoost, and multilayer neural networks. These comparisons help determine whether any improvement arises from nonlinear representation learning, the boosting structure, or the treatment of class imbalance.

3. Data Sources

The primary data source considered in the project is historical and synthetic bank transaction data available through the Smart Data Foundry (SDF). In addition to the SDF data, five publicly available datasets have been identified as sources for initial implementation and external benchmarking. These datasets differ in their application, unit of analysis, and definition of the minority outcome. They are therefore treated as methodological benchmarks rather than direct substitutes for the SDF transaction data.

- **Bank Marketing**: 36,168 observations, 17 features, and an 11.61% minority class. The outcome represents customer responses rather than confirmed fraud. The dataset is therefore relevant only as a general benchmark for rare-event classification.
  https://archive.ics.uci.edu/ml/datasets/bank+marketing
- **Vehicle Insurance**: 11,564 observations, 32 features, and a 5.98% minority class. The dataset contains information on potentially fraudulent vehicle-insurance claims.
  https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection
- **Fraudulent Claims on Cars**: 17,998 observations, 26 features, and a 15.65% minority class. The dataset covers suspected fraud associated with automobile physical-damage insurance claims.
  https://www.kaggle.com/datasets/surekharamireddy/fraudulent-claim-on-cars-physical-damage
- **Worldline and ULB Credit Card Fraud**: 284,807 observations, 29 model features, and a fraud rate of approximately 0.17%. Its extreme class imbalance makes it a relevant benchmark for transaction-level fraud detection.
  https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
- **BankSim**: 594,643 simulated payment transactions, 5 model features, and a fraud rate of approximately 1.21%. Its large sample size makes it useful for examining rare-event detection and computational scalability.
  https://www.kaggle.com/datasets/ealaxi/banksim1

As an initial validation exercise, we have used the BankSim dataset to implement a transaction-level fraud-detection workflow. Traditional machine-learning models have been trained and evaluated to distinguish fraudulent transactions from legitimate transactions. These experiments provide initial

benchmark results and demonstrate the feasibility of applying the modelling workflow to a large and highly imbalanced transaction dataset.

4. Project Outcomes

The research has four related outcomes. First, it establishes an analytical framework for identifying behavioural, financial, and transaction-sequence characteristics associated with vulnerability to fraud. These characteristics support the construction of customer profiles and the segmentation of individuals according to estimated levels of risk. Second, the project develops an interpretable early-warning model that assigns risk scores to customers, transactions, or both, depending on the available outcome labels. The framework combines nonlinear representation learning with a decision structure that can be examined by researchers and practitioners. Third, the resulting customer profiles and risk scores inform the assessment of practical protection mechanisms. Potential applications include personalised fraud alerts, adaptive transaction thresholds, enhanced monitoring, customer risk segmentation, and dynamic intervention strategies. These mechanisms allow financial institutions to direct additional protection towards customers or transactions associated with elevated risk while limiting unnecessary disruption to legitimate activity. Fourth, the empirical analysis assesses whether the framework operates effectively in high-volume transaction environments. This assessment considers predictive performance, false-positive rates, computational requirements, stability across different samples, and model interpretability. Comparisons with established benchmark models determine whether the framework provides incremental predictive value. Overall, the project establishes a research design for examining whether advanced but interpretable machine-learning methods can identify fraud vulnerability and support earlier and more targeted intervention.

References

Boulieris, P., Pavlopoulos, J., Xenos, A., and Vassalos, V. (2024). Fraud detection with natural language processing. *Machine Learning*, 113(8):5087–5108.

Fanai, H. and Abbasimehr, H. (2023). A novel combined approach based on deep autoencoder and deep classifiers for credit card fraud detection. *Expert Systems with Applications*, 217:119562.

Forough, J. and Momtazi, S. (2021). Ensemble of deep sequential models for credit card fraud detection. *Applied Soft Computing*, 99:106883.

Motie, S. and Raahemi, B. (2024). Financial fraud detection using graph neural networks: A systematic review. *Expert Systems with Applications*, 240:122156.

Xu, B., Wang, Y., Liao, X., and Wang, K. (2023). Efficient fraud detection using deep boosting decision trees. *Decision Support Systems*, 175:114037.